# The EU DataGrid Tutorial

The EU DataGrid Tutorial Team
`http:cern.ch/edgtutor`
Editors: Erwin Laure, Heinz & Kurt Stockinger*

**Abstract**

The EU DataGrid project (EDG) [1] is not only a software provider of Grid software but also puts much emphasis on training. For this purpose, a tutorial programme has been created and successfully presented at several events all over the world.

The tutorial covers various aspects of Grid computing and provides the ability to get hands-on experience with modern Grid tools. A major part is dedicated to Grid software that has been produced within the EDG project who's aim is to develop high level Grid middleware and to operate a large-scale research testbed for Grid computing. The tutorial presents the EDG software architecture and discusses the interplay of the basic Grid software (Globus, CondorG), higher level EDG middleware, and application software on the EDG testbed. Emphasis is put on specific middleware issues in job submission, data management and information systems as well as on EDG's security architecture. In several exercises students learn how to use Grid tools for their distributed data or computing intensive applications.

## 1 Introduction

In the past few years, many Grid projects worldwide have developed grid solutions which go beyond a simple proof of concept and allow the exploitation of Grid computing, i.e. world-wide resource sharing within specific communities, so-called *Virtual Organisations (VOs)*, in an ever growing scale. However, Grid computing is still not in the mainstream although several communities have already adopted this technology as main production infrastructure.

This is mainly due to the relative immaturity and complexity of Grid software which requires specific skills and experience in dealing with these tools in order to efficiently exploit them. Although work continues in stabilising Grid software and developing higher level Grid tools providing better usability, it is essential that a wider user community is attracted by Grid computing already in this stage to drive the further development

*Contact: `Heinz.Stockinger@cern.ch`

addressing their specific needs. We believe that promoting Grid computing requires a substantial training effort in order to attract and train potential users, developers, managers and interested people.

The EDG project is one of the major providers of Grid software and has a large user community all over the globe. The project is in the final phase and a testbed spanning some ten major sites all over Europe has been up and running since the beginning of 2002. Three application domains are using this testbed to explore the potential that Grid computing has for their production environments: Particle physics, Earth observation and Biomedics.

As part of the training and dissemination effort a tutorial programme has been created that covers general aspects of Grid computing as well as the main parts of the EDG software system. It is mainly presented from a user's point of view but it also gives insights for developers and some hints for system administrators. The main aim is to attract and train new users but also give software developers an overview of the different components within the EDG Grid middleware. Specific tracks on installing and running Grid middleware, targeted towards system administrators of Grid sites, are also available, but their description is beyond the scope of this paper.

A conventional EDG tutorial consists of an 8 hours lecture programme as well as 8 hours practical hands-on exercises where students have access to a Grid testbed. The goal is to provide access to the software that is taught in the lectures.

In the remainder of this paper we give an overview about the main topics covered in the tutorial, discuss the hand-on exercises which are an integrated part of the tutorial and provide information on how to obtain the tutorial material and the pre-requisites required to run a tutorial.

## 2 Tutorial Programme

The tutorial program consists of nine lectures which introduce the students in Grid computing in general, discuss the EDG software components and their deployment on the EDG testbed, show examples on how application groups successfully exploit Grid computing for solving their problems, and finally give an outlook on future directions of Grid computing:

1. **Introduction to Grid Computing & EU DataGrid Project**: An overview is given about Grid computing in general and the EDG project. The project organisation and the general software architecture are described. In addition, related projects are outlined.

   *Key items*: Grid Computing, Data Grids, international Grid projects world-wide, EU DataGrid project

2. **Security Issues**: Secure access to Grid resources is a major issue and one of the first things a user has to deal with when starting to use the Grid. A basic overview about current security solutions in the EDG project is given.

*Key items*: GSI Security (*Grid Security Infrastructure* provided by the globus project), user and host certificates, Virtual Organisation Management

3. **Testbed Overview**: EDG deploys a large-scale testbed that spans several sites all over Europe. Definitions are given about what Grid services and resources are available and where they can be used. A detailed overview about the testbed is given.

   *Key items*: logical machine types (User Interface, Storage Element, Computing Element, Worker Node, Information Service, Resource Broker, etc.), overview about EDG's international testbed

4. **Workload Management**: Most of EDG users interact with the EDG software system by submitting their jobs (executable programs) to a Resource Broker which does a matchmaking on available and requested resources and then dispatches jobs to resources in the testbed. This lecture provides background about the work load management software system and details about job submission.

   *Key items*: user interaction with Workload Management System (WMS), components (Resource Broker, Logging & Bookkeeping, etc.), Job Description Language (JDL)

5. **Data Management**: One of the main objectives of a Data Grid is the management of large distributed data stores. This lecture gives an overview about replica and meta data management as well as the software tools provided by EDG to deal with data management problems.

   *Key items*: replica management system, Replica Location Service (including Replica Metadata Catalogue), Replica Access Optimisation, Storage Resource Management

6. **Information Service**: In a Grid environment, there are several hardware and software resources that can be used by end-users as well as Grid services and applications. Information systems are used to keep track of resources and also to monitor the current status. The EDG solution is outlined in detail and how end-users can interact with it.

   *Key items*: Relational Grid Monitoring Architecture (R-GMA), Consumer, Producer, Registry, Archiver, Glue Schema

7. **Software Installation/Configuration**: This lecture gives a brief introduction on how to obtain EDG software and how to installation and configuration of EDG software tools.

   *Key items*: LCFG, EDG software repository

8. **Applications**: In the EDG project, three major application domains are supported: High Energy Physics, Earth Observation and Biomedical Applications. The talk gives a brief overview about these applications and how they use Grid tools.

9. **Future Direction**: This lecture briefly covers the future of the EU DataGrid project as well as Grid computing in general.

   *Key items*: SOAP, OGSA, Grid Services, Web Services

Each of the lectures is between 30 and 45 minutes long and in a typical setup lectures 1-4 are given in the first day, and lectures 5-9 during the second day. Lectures are typically followed by hands-on exercises which allow the students to get real experience with the topics covered in the lectures. The following section discusses these exercises in more detail.

# 3 Hands-on Exercises

One of the main goals of the tutorial programme is to give students hands-on experience with the EDG software on a distributed Grid testbed. Usually, the Grid Dissemination Testbed (GriDis) [3] is used for that purpose. GriDis hosts the main Grid infrastructure and relies on the computing resources available in the EDG testbed. In order to increase the testbed resources, sites from other projects, in particular from the EU CrossGrid [4] project, can be temporarily added to the testbed available to the students.

The exercises are typically performed on client machines or laptops from where the students log in to the testbed's *User Interface*, the gateway to the Grid which hosts all the client software required to interact with the Grid.

In the hands-on session we focus on the following three aspects: job submission, data management and information systems. For each of these areas, students are given exercises with the respective solutions using both command line tools and C++ or Java APIs. This allows the students the get experience with real usage of a Grid environment.

# 4 Tutorial Material & Website

The main source for tutorial material is the Tutorial website [2] where one can find links to the lecture slides and all the handout material provided to students in the hands-on session. The web page is also the main point of communication during the hands-on session since it is always up-to-date with the latest information on the testbed and software versions to use. All agendas of future and past tutorials are linked on the web page, too.

Institutions wishing to host a DataGrid tutorial need to provide the infrastructure for giving the lectures and the local infrastructure to allow the students to connect to the Grid. In particular, a data projector is required for the lectures, and apart from the students terminals, a minimum of two machines with high bandwidth connection to the Internet are required for the hands-on exercises. These machines, which need to run GNU/Linux RedHat 7.3, are configured as User Interfaces, i.e. they host the Grid clients. Typically, one User Interface is shared among about 20 students.

# 5 Conclusion

Several hundreds of people have been trained already in more than a dozen tutorials all over the world. Based on the students' feedback we constantly improve the tutorial material and thus provide a good training infrastructure for people interested in Grid computing.

## Acknowledgements

Several people from the EDG and LCG projects have contributed to make the EDG tutorials a success. Thanks to all of you for your good team spirit.

This work was partially funded by the European Commission programme IST-2000-25182 through the EU DataGrid Project.

# References

[1] EU DataGrid project (EDG): http://ww-eu-datagrid.org

[2] EDG Tutorial web site: http://cern.ch/edgtutor

[3] Grid Dissemination Testbed (GriDis):
http://web.datagrid.cnr.it/GriDis/GriDisWP1.html

[4] EU CrossGrid project: http://www.eu-crossgrid.org